miRNAMatcher: High throughput miRNA discovery using regular expressions obtained via a genetic algorithm.

Eugene Duvenage

A thesis submitted in fulfilment of the requirements for the degree of

Magister Scientiae in Bioinformatics at the South African National Bioinformatics

Institute, Faculty of Science, University of the Western Cape

Supervisor: Prof. V. Bajic

October 2008

Keywords

miRNA

Gene expression regulation

Computational miRNA identification

Hairpin structural motifs

Secondary structure calculation

Machine learning

Genetic algorithm

Regular expressions

Genome scan

High throughput

Abstract

miRNAMatcher: High throughput miRNA discovery using regular expressions obtained via a genetic algorithm.

E. Duvenage - MSc. thesis, South African National Bioinformatics Institute, Faculty of Science, University of the Western Cape.

Micro-RNA (miRNA) are short, approximately 22 nucleotide long, non-coding RNA molecules that are involved in the regulation of gene expression. miRNAMatcher is a software program developed to scan genomic DNA in order to search for miRNA candidates. miRNAMatcher performs this task by using regular expressions to match structural motifs that have been discovered by employing a genetic algorithm machine learning technique. The genetic algorithm identifies sets of short 5-7 base motifs that uniquely identify a particular miRNA family and converts the motifs and their relative positions to a simple regular expression. Consensus classifiers are built from three sets of regular expressions per miRNA family. miRNAMatcher was used to scan chromosomes one, two and eight of *canis-familiaris* which took approximately seventy four hours per chromosome scanning up to two hundred and fifty million bases per chromosome. For chromosomes one and two all known miRNA were identified, with just over fifty percent found for chromosome eight. A number of exact and near exact matches to miRNA known in other species were also identified. Many unknown miRNA candidates were also suggested by the results. miRNAMatcher is algorithmically simple and efficient and does not rely on computationally intensive secondary structure calculations employed by many other miRNA discovery tools. The resultant regular expressions are portable across many different software platforms making it an attractive option for miRNA discovery. A number of improvements for future versions have been identified by this study.

October 2008

Declaration

I declare " **miRNAMatcher: High throughput miRNA discovery using regular expressions obtained via a genetic algorithm** " to be my work, which has not been submitted for any degree or examination in any other institution, and that all the sources I have used or quoted have been indicated and acknowledged by complete references.

Full Name: Eugene Duvenage Date: 17 October 2008

Signed: _____

Table of Contents

1. Introduction

In order to understand the difficulties involved in the computational discovery of micro-RNA (miRNA), the basics of miRNA biology will be discussed using Figure 1 to illustrate the process. miRNA consist of approximately 22 nucleotide long non-coding RNAs which function as gene expression regulators (Clarke & Sanseau, 2007). In animals, single stranded miRNA, hundreds to thousands of nucleotides in length, are transcribed in the nucleus by polymerase II as primary miRNA (pri-miRNA) transcripts (Lee, et al., 2004). These transcripts are then processed by Drosha (Lee, et al., 2003) (Zeng & Cullen, 2005), a ribonuclease III enzyme, into approximately seventy nucleotide long precursor miRNA (pre-miRNA), that form a distinctive hairpin secondary structure. The pre-miRNA is then exported from the nucleus to the cell cytoplasm by the Exportin 5 protein which recognises the dangling end 3' overhang of the hairpin created by the ribonuclease III cleavage (Yi, Qin, Macara, & Cullen, 2003)(Bohnsack, Czaplinski, & Görlich, 2004). The ribonuclease III enzyme Dicer then cleaves the pre-miRNA at the hairpin end to leave a double stranded miRNA (Bernstein, Caudy, Hammond, & Hannon, 2001). This double stranded miRNA duplex is then assembled into the RNA induced silencing complex (RISC) as a mature miRNA with the degradation of the complimentary strand (Clarke & Sanseau, 2007). The activated RISC complex then either regulates gene expression via site specific cleavage when the miRNA has perfect sequence complimentarity with a target messenger RNA site or otherwise via translational inhibition when there is imperfect sequence complimentarity (Gregory, Chendrimada, Cooch, & Shiekhattar, 2005), (Filipowicz, 2005), (Chendrimada, et al., 2007).

From the basic biology above a number of possible features can be identified which can be used when trying to identify miRNA with computational techniques. Examples, some seen in the miRNA shown in Figure 2, would be the distinctive stem-loop structure, Drosha and

Dicer interaction and the fact that miRNA are often strongly conserved across species (Berezikov, Guryev, van de Belt, Wienholds, Plasterk, & Cuppen, 2005), (Bentwich, et al., 2005), although there is evidence to support the fact that there are many miRNA that are not highly conserved (Bentwich, et al., 2005).

There are 6396 known miRNA in version 11 of the Sanger miRBase registry (April 2008) (Griffiths-Jones, Grocock, van Dongen, Bateman, & Enright, 2006), (Griffiths-Jones, Saini, van Dongen, & Enright, 2008), with more than 650 of those found in the human genome. Considering that the human genome has over 3,000,000,000 base pairs according to the NCBI (version 36) human genome assembly, identifying an approximately 70 base long miRNA is no easy task.

Figure 1 miRNA Biogenesis

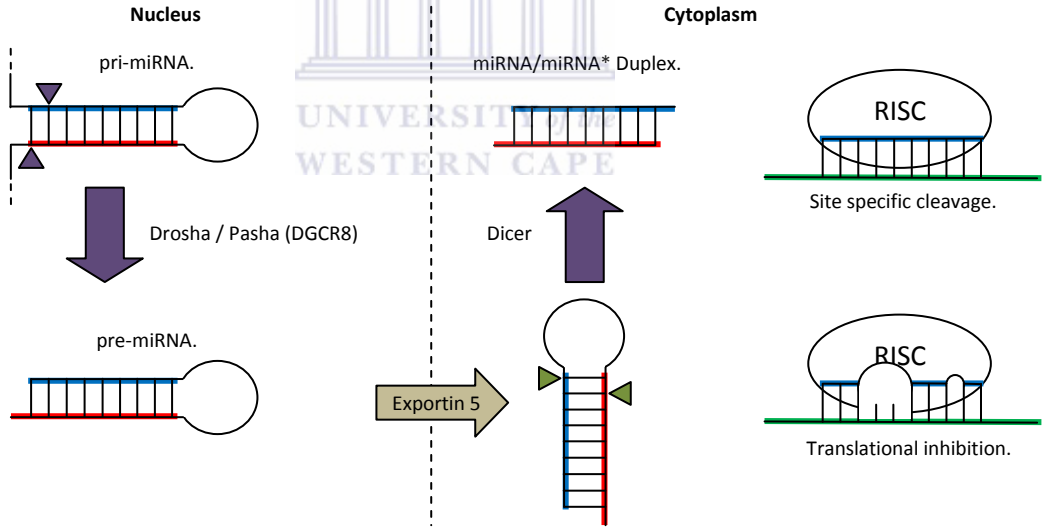
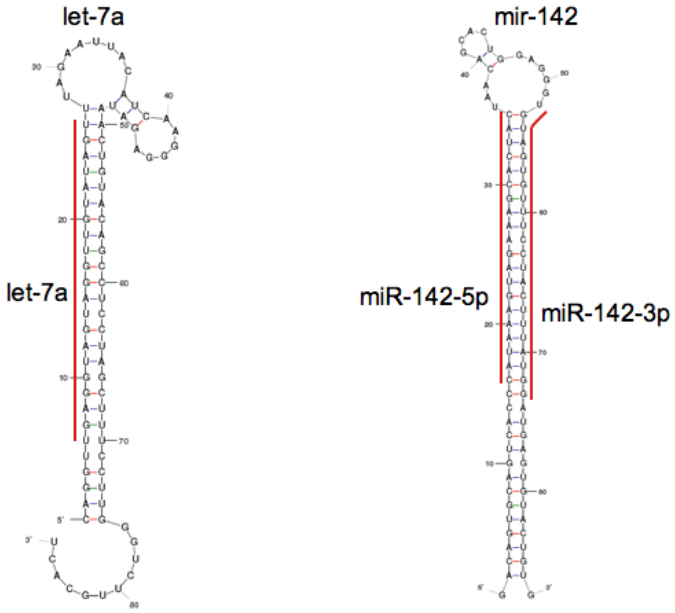


Figure 2 Example of miRNA secondary structure with prominent stem loop structures.



2. Problem Definition

There are currently two major methods used to identify miRNA from genomic sequence data (Hertel & Stadler, 2006). The first uses sequence homology to find sequence segments across many species that have similar sequence structure when compared to an experimentally discovered miRNA. A tool which uses this method is miRNAminer (Artzi, Kiezun, & Shomron, 2008), it uses BLAST (NCBI BLAST) to find matches in target genomes using the pre-miRNA as the input sequence, further filters are then applied including calculating the RNA secondary-structure folding energy using RNAFold (RNA Secondary Structure Prediction and Comparison), requiring a hairpin shape secondary structure and further mature miRNA sequence conservation statistics. Using the BLAST algorithm (Altschul, Gish, Miller, Myers, & Lipman, 1990) requires finding approximate local sequence alignments for the input miRNA with all possible target DNA sequences in the BLAST database. While this is a highly optimised algorithm it is still computationally expensive which consumes more computation time as more DNA sequence data is added to the database.

The second technique does not use sequence homology and concentrates on using machine learning techniques to identify miRNA using predominantly structural features. Machine learning techniques can be described in a simple manner as the automated acquisition of knowledge from a set of known observables. During the training cycle features are extracted and mapped to known outcomes to construct classifiers. This "automated" learning technique can only be successful when sufficient training data is available. It is common to place a sliding window over genomic DNA to produce a sequence segment from which features are extracted. Examples of such features would be the secondary structure minimum free energy using RNAFold (RNA Secondary Structure Prediction and Comparison), whether a hairpin loop exists, the % of GC base paring and

others that are then fed as input to the classifier. Software programs such as mirCoS (Sheng, Engstrom, & Lenhard, 2007) use a support vector machine (SVM) learning technique to filter once the folding energy calculations have produced a set of possible candidates. While the latter technique is not constrained by requiring an existing miRNA starting point, machine learning techniques require plentiful and relevant data to arrive at a good model of the problem. The version 11 Sanger miRBase registry (April 2008) (Griffiths-Jones, Grocock, van Dongen, Bateman, & Enright, 2006), (Griffiths-Jones, Saini, van Dongen, & Enright, 2008) contains a little over 550 miRNA families, only 134 families have at least 10 members, making it difficult to have high confidence in classifiers trained on poorly represented families.

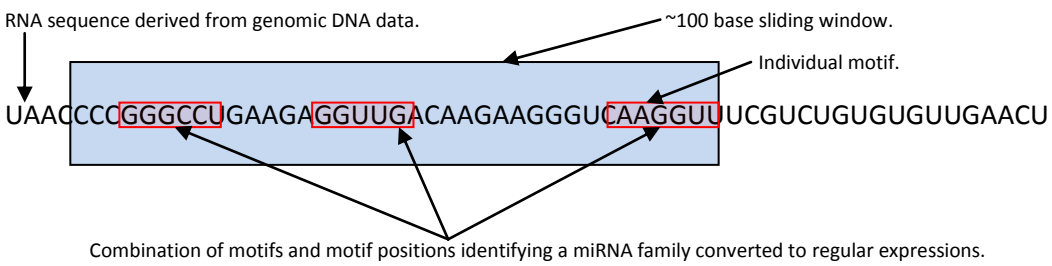
In summary there currently exist techniques to discover miRNA however both require many calculations to be performed during the identification limiting their use at a genomic level. Machine learning techniques are currently providing the best results by combining a number of calculated and statistically derived features to identify miRNA candidates, however almost all of these still include computationally intensive secondary-structure calculations. It is the aim of this project to produce a miRNA identification process that minimises and simplifies the number of computational elements required during the identification process.

3. Strategy

The strategy employed aims to avoid the use of computationally expensive secondary-structure calculations, instead only simple and efficient text searching techniques, namely regular expressions, should be used to search genomic DNA sequence data to identify structural motifs. Identifying these motifs in a DNA sequence segment will indicate that the region has structural commonality with miRNA from a particular miRNA family. A machine learning technique, specifically a genetic algorithm, will be employed to find the set of motifs that will allow each miRNA family to be characterised independently, such that the resulting regular expressions accurately identify each miRNA family as best as possible. While the training of the algorithm will be computationally intensive, the outcome is merely a set of regular expressions that can be used from a multitude of software environments to search for miRNA in a relatively efficient manner.

The genetic algorithm will try to find, for each known miRNA family in version 11 of the Sanger miRBase registry (April 2008) (Griffiths-Jones, Grocock, van Dongen, Bateman, & Enright, 2006), (Griffiths-Jones, Saini, van Dongen, & Enright, 2008), the best combination of short (5-7 base long) motifs as well as the distances between them that will uniquely identify each miRNA family.

Figure 3 Motif identification and combination



4. Algorithm

4.1.Training

The first part of the algorithm deals with the definition of various aspects of the genetic algorithm such as the functional chromosome encoding, fitness function and the training. Genetic algorithms are a form of evolutionary algorithm in which a population of encoded potential solutions evolve and are optimised towards the best solution for the domain it is being applied towards. Before describing how this particular genetic algorithm encodes the problem of identifying miRNA the available input data needs to be discussed.

4.1.1. Data

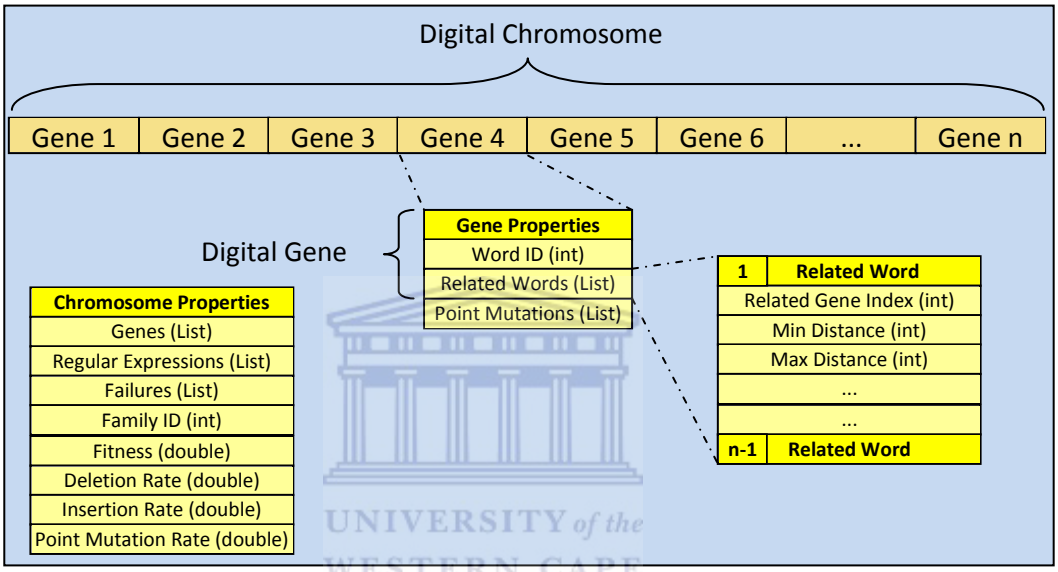
Version 11 of the Sanger miRBase registry (April 2008) (Griffiths-Jones, Grocock, van Dongen, Bateman, & Enright, 2006), (Griffiths-Jones, Saini, van Dongen, & Enright, 2008) provides a list of 6396 known miRNA categorised into a little over 550 families with a number of miRNA not yet assigned to a family. The data preparation began by separating out those miRNA that had been categorised into a family and belong to species contained in the ENSEMBL(Birney, Andrews, & Bevan, 2004) genome database. The reason for requiring the miRNA to belong to a species in the ENSEMBL (Birney, Andrews, & Bevan, 2004) genome database is so that the miRNA sequence given in the Sanger miRBase registry could be extended to at least 100 base pairs, resulting in around 4000 available miRNA. A length of 100 bases is chosen so that the sequence segment includes the full pre-miRNA ,(70-90 nucleotides in length), and short flanking regions on either side. We then randomly separated this filtered dataset into a test and training set on a per family basis so that roughly the same number of miRNA of each miRNA family was represented in each dataset. Only the training dataset was used for the rest of the training process including any data manipulation and generation of statistics. For each miRNA belonging to a particular miRNA family all unique five, six and seven base motifs as well as the distances

between each same length motif were identified for the extended miRNA sequence area (100 bases) and stored in a database. Motifs less than 5 bases in length produced too many motifs to easily compute the distance matrices while motifs above 7 produced too few useable combinations for some miRNA families. For each family the top 100 motifs of each size (five, six and seven bases) that appeared most often in the extended miRNA sequences were found across all members of that miRNA family again storing the results in a database. These top 100 motifs were then further analysed to find the minimum and maximum distances between the motifs for all miRNA in each family and stored in a database thus creating a single combined min/max motif distance matrix for each family. While every miRNA that belonged to a family was used in the initial separation of training and test data some miRNA families do not have many members. The machine learning techniques for training of the classifiers may yield poor results for these cases and as such we will limit our discussion of the results to families containing at least 8 members.

4.1.2. Genetic algorithm encoding

Genetic algorithms use the principals of biological evolution to create digital populations of genes and chromosomes that undergo genetic operations such as mutation, insertion, deletion and natural selection of the fittest candidates for a specific function. A genetic algorithm was setup to build a classifier for each miRNA family.

Figure 4 Genetic algorithm functional encoding



The genes of the digital chromosome hold an integer id of a motif belonging to the miRNA family, a list holding relations to all other genes in that particular chromosome and a list of possible point mutation locations. The related motifs in the list each hold an integer index of the gene it represents in the chromosome and the minimum and maximum distances found when analysing the distance between the motifs in the data preparation stage. A minimum and maximum length is specified for the chromosomes when configuring the training session, the insertion and deletion genetic operations could alter a particular chromosome length during a training session between these values.

4.1.3. Genetic Operators

The genetic algorithm framework mimics the biological genetic operators found in nature and with each generation applies operations such as insertion, deletion, crossover and point mutations. The chance of any of these operators occurring is controlled via the corresponding rate variables specified when configuring the training session. While the operations described below are quite simple the actual implementation is rather more robust such that invalid encodings of motifs do not occur.

- **Cross-over –** randomly determined sections of two chromosomes are crossed over, effectively swapping a portion of the motifs encoded in the participants.
- **Insertion –** a further gene is inserted into the digital chromosome representing a motif from the miRNA family that has not yet been used in the particular chromosome.
- **Deletion –** a gene is removed from the chromosome, effectively removing a motif.
- **Point mutations –** if a motif is selected to allow a point mutation, a position in the motif for the point mutation is randomly selected. This location will then be allowed to take on any of the 4 possible RNA bases (A, G, C, or U). More than one point location mutation is allowed on a motif however the possibility of multiple point mutations has been made significantly lower than a single point mutation.

4.1.4. Genetic algorithm fitness function

The fitness function of a genetic algorithm determines how members of a population of digital chromosomes are ranked with regards to their ability to solve a problem. In this case the goal is to find a set of motifs that can be converted into regular expressions such that when applied to an RNA sequence a miRNA belonging to the appropriate miRNA family can be identified.

At the start of the genetic algorithm training session for a single miRNA family a population of 500 chromosomes are randomly populated with motifs for the miRNA family found earlier during the data processing stage. For each generation of the training each chromosome is ranked by the fitness function with the best candidates being used as parents for the next generation. The number of generations used in training sessions was between 25 and 250.

The actual ranking occurs in a number of stages, a chromosome is first converted to a set of regular expressions. The conversion to a set of regular expressions is straight forward as each gene represents a motif that can be converted to a string, such as **AGCUA** for a 5 base motif. Each gene also has a list of motifs that it relates to as well as the minimum and maximum distance between the motifs, this can be used to build a regular expression such as **AGCUA[ACGU]{3,7}AAAGG.** This regular expression matches the motif **AGCUA** then allows any combination of bases (A, C, G or U) between 3 and 7 bases in length and then matches the second motif **AAAGG**. If a gene has been allowed a point mutation then this can be represented in the example above as **AGCUA[ACGU]{3,7}(A (?:A|C|G|U) AGG)**, which matches the motif **AGCUA** then allows any combination of bases (A, C, G or U) between 3 and 7 bases in length and then matches the first letter of the motif **A**, followed by the point mutation, any single base (A, C, G or U) followed by the rest of the motif **AGG**. Point mutations can be turned off completely as enabling them impacts training

performance severely. We also introduce a “wobble factor” a tolerance amount that is added to minimum and maximum distance measures. This value is aimed at managing the drift of motifs over time and species due to insertions and deletions. Values from 0 to 8 bases were used during training sessions. For some miRNA families, particularly when using the shorter length motifs several motifs tend to overlap producing a regular expression that consists of the combined motifs with no defined distance gap. For example the overlapping motifs **AGUGA** and **UGAUA** combine to form **AGUGAUA**. These form of overlaps result in too many false positives as they may occur randomly in the genomic data, they are therefore discouraged by negatively influencing their rank. The aim of the regular expressions generated from the digital chromosome is thus to provide a number of motifs (between 3 and 10), either as exact sequence segments or with some variation (via point mutations), that have a predefined range of distances between each motif that uniquely identifies the miRNA family. All unique combinations for the chromosome are generated as regular expressions and stored.

The second phase of the fitness function is the actual evaluation of each chromosome for ranking purposes. The training dataset which contains both miRNA sequences that belong to the chromosomes miRNA family and many more that do not is used. For each 100 base miRNA segment in the training dataset, the chromosomes set of regular expressions is applied and it is recorded how many of the regular expressions produce a match as well as how many mismatches occur. A match is recorded if a certain amount of the chromosomes regular expressions produces a positive result when applied and the miRNA being tested belongs to the correct miRNA family. The number of allowed regular expression failures was set at one and three with three giving the best results. A mismatch occurs if a match occurs when the miRNA being tested does not belong to the correct miRNA family. A tolerance of 1 percent has been allowed with regards to mismatches, with any higher mismatch rate resulting in a chromosome immediately receiving a rank of zero. Requiring a

mismatch rate of zero leaves a small amount of families with no classifier at all, other techniques will be introduced in later sections detailing how multiple matches from different classifiers are handled. The ranking of matches and mismatches are converted to a single numerical ranking in the following manner:

```
double Ranking =
( ( (double)RegExMatched.Count / (double)FamilyMemberCount ) +
( 1.0 - (double)(NonMatched.Count) / (double)TrainingMembers.Count
)) / 2.0 * 100.0;
```

The pink area above correponds to the chromosomes ability to correctly classifiy the miRNA family members. The light blue area is a measure of the amount of failures, whether false positives or false negatives, these are combined such that when all family members are classified and no mismatches occur a vaule of 100% is reached.

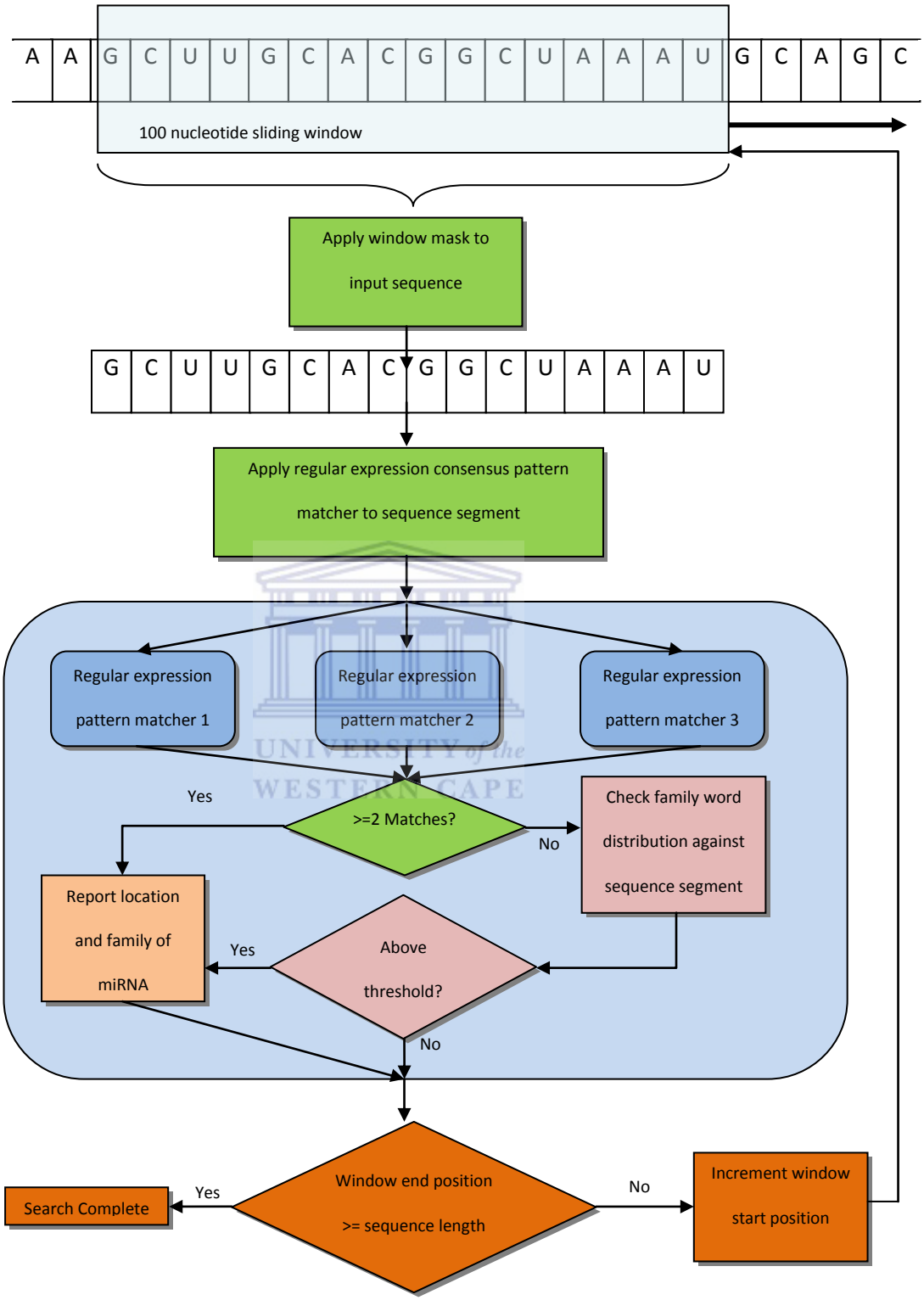
The fitness function also contains three types of optimisers for the classification problem. The first minimises the number of words in the chromosome, the second maximises the number of words in the chromosome and the third randomly selects whether a longer or shorter chromosome is preferred during selection. The reason for the inclusion of a number of variable paramters is to facilitate multiple training sessions with different configurations to produce a variety of good results. The more varied the regular expressions generated are the easier it is to build a consensus classifier out of the three best genetic algorithm training sessions. The variables that were altered during training runs were firstly the motif length, for each motif length training sessions included runs with and without point mutations, runs with a small wobble factor (0-3) and runs with a large wobble factor (4-8) all combined with the 3 optimisations functions (minimising, maximising and randomly deciding on chromosome length). The top three classifiers were picked to be the ones that correctly matched the highest number of unseen miRNA from

the unused test dataset to their families. The hope is that if a family has some significant partition of motifs amoungst its members that all the partitions will be covered by at least one set of training.

4.2.Usage

The algorithm used when scanning genomic data to identify miRNA is based on the use of the three consensus classifiers built during the training phase. The miRNAMatcher software was created using the c# programming language, which builds on the Mono project runtime for Linux or Microsoft Windows users, as a command line application as well as a re-useable class library. The program slides a 100 nucleotide window over the input genomic data and for each window segment applies the regular expression sets from the top three training sessions for each miRNA family. The program searches both the sense and anti-sense RNA strands. For each family if two or more classifiers report a match a match is immediately given. If only one classifier reports a match the match is further investigated by counting the number of motifs in the segment that belong to the miRNA family's training data, if the figure is below the minimum for all members of the training data the match is disregarded, otherwise it is recorded as a match. If more than one family is matched the matches are ordered by their score, i.e. if only two of the three classifiers matched in one family but all three in another family matched then the highest ranking classifier is reported as a match. If a tie is reached a match and candidate families are reported but no decision on the family is made.

Figure 5 Regular expression based miRNA discovery algorithm



5. Results

5.1. Tests on identifying known miRNA

In order to test the abilities of miRNAMatcher comparative tests were run against two previously published software programs designed to identify miRNA namely miR-abela (miRNA prediction with miR-abela) and PROMIR II (Kim, Kim, Kim, & Zhang, 2006). Both of these programs are available to use via a web interface, miR-abela can be found at http://www.mirz.unibas.ch/cgi/pred_miRNA_genes.cgi and PROMIR II can be found at http://cbit.snu.ac.kr/~ProMiR2/. miR-abela is based on a support vector machine learning technique as described in chapter 2 while PROMiR2 uses a combination of the sequence homology and machine learning techniques with several input configurations available that combine the techniques. Wrapper code was written for both web sites that allowed all 3 programs to be instantiated in the same manner such that each program could be passed a segment of RNA sequence data and a true or false result could be obtained with regards to the segment containing an miRNA or not. miRNA from ten species contained in the ENSEMBL (Birney, Andrews, & Bevan, 2004) database were chosen from the unseen test dataset that miRNAMatcher did not use during training, it was not established whether the other programs had used any of the data during their creation. The ten species chosen from the unseen test data were:

Table 1 List of species and known miRNA counts for the unseen test dataset

Species	No. miRNA
Human (*homo sapiens*)	130
Mouse (*mus musculus*)	111
Rat (*rattus norvegicus*)	79
Chicken (*gallus gallus*)	56
Rhesus Macaque (*macaca mullata*)	101
Dog (*canis familiaris*)	62
Chimp (*pan troglodytes*)	41
Cow (*bos taurus*)	39
Opossum (*monodelphis domestica*)	50

Zebra Fish (*danio rerio*)	125

While miR-abela does not require an organism to be chosen as part of its input parameters PROMIR II was used in its –g (generalised) mode and only supported four of the chosen species. All individual parameters were left as set by default on the websites.

miRNA were also chosen from the training set as follows:

Table 2 List of species and known miRNA counts for the training dataset

Species	No. miRNA
Human (*homo sapiens*)	140
Mouse (*mus musculus*)	112
Rat (*rattus norvegicus*)	72
Chicken (*gallus gallus*)	54
Rhesus Macaque (*macaca mullata*)	96
Dog (*canis familiaris*)	64
Chimp (*pan troglodytes*)	29
Cow (*bos taurus*)	46
Opossum (*monodelphis domestica*)	29
Zebra Fish (*danio rerio*)	98

miR-abela and PROMIR II have slightly different designs to miRNAMatcher in that they have no concept of a miRNA family and just report on whether an miRNA is found, miRNAMatcher reports on whether a region of around 100 bases has been found containing a miRNA of a particular family and only correctly identified families were counted as matches. As miR-abela and PROMIR II were accessed via a web interface no statistics can be gathered on their actual performance which took in the order of low tens of minutes per species, miRNAMatcher took under five seconds to classify all miRNA in a particular species for the test data. Table 3 summarises the results of each classifier when run against the unseen dataset while Table 4 does the same for the training dataset. Table 5 summarises the results for the combined training and unseen testing datasets.

Table 3 Results for miRNAMatcher, miR-abela and PROMIR II for the unseen dataset

Unseen Test Dataset

Tool		miRNAMatcher	
Species	**Total miRNA**		
Human (*homo sapiens*)	130	124	95.38%
Mouse (*mus musculus*)	111	104	93.69%
Rat (*rattus norvegicus*)	79	75	94.94%
Chicken (*gallus gallus*)	56	56	100.00%
Rhesus Macaque (*macaca mullata*)	101	94	93.07%
Dog (*canis familiaris*)	62	60	96.77%
Chimp (*pan troglodytes*)	41	39	95.12%
Cow (*bos taurus*)	39	39	100.00%
Opossum (*monodelphis domestica*)	50	45	90.00%
Zebra Fish (*danio rerio*)	125	103	82.40%
	794	**739**	**93.07%**

Unseen Test Dataset

Tool		miR-abela	
Species	**Total miRNA**		
Human (*homo sapiens*)	130	109	83.85%
Mouse (*mus musculus*)	111	91	81.98%
Rat (*rattus norvegicus*)	79	59	74.68%
Chicken (*gallus gallus*)	56	45	80.36%
Rhesus Macaque (*macaca mullata*)	101	81	80.20%
Dog (*canis familiaris*)	62	55	88.71%
Chimp (*pan troglodytes*)	41	27	65.85%
Cow (*bos taurus*)	39	30	76.92%
Opossum (*monodelphis domestica*)	50	35	70.00%
Zebra Fish (*danio rerio*)	125	100	80.00%
	794	**632**	**79.60%**

Unseen Test Dataset

Tool		PROMIR II	
Species	**Total miRNA**		
Human (*homo sapiens*)	130	96	73.85%
Mouse (*mus musculus*)	111	64	57.66%
Rat (*rattus norvegicus*)	79	45	56.96%
Chicken (*gallus gallus*)	56	46	82.14%
Rhesus Macaque (*macaca mullata*)	101	N/A	0.00%
Dog (*canis familiaris*)	62	N/A	0.00%
Chimp (*pan troglodytes*)	41	N/A	0.00%
Cow (*bos taurus*)	39	N/A	0.00%
Opossum (*monodelphis domestica*)	50	N/A	0.00%
Zebra Fish (*danio rerio*)	125	N/A	0.00%
	794(376)	**251**	**66.76%**

Table 4 Results for miRNAMatcher, miR-abela and PROMIR II for the training dataset

Training Dataset

Tool		miRNAMatcher	
Species	**Total miRNA**		
Human (*homo sapiens*)	140	130	92.86%
Mouse (*mus musculus*)	112	109	97.32%
Rat (*rattus norvegicus*)	72	71	98.61%
Chicken (*gallus gallus*)	54	54	100.00%
Rhesus Macaque (*macaca mullata*)	96	94	97.92%
Dog (*canis familiaris*)	64	62	96.88%
Chimp (*pan troglodytes*)	29	29	100.00%
Cow (*bos taurus*)	46	46	100.00%
Opossum (*monodelphis domestica*)	29	27	93.10%
Zebra Fish (*danio rerio*)	98	79	80.61%
	740	**701**	**94.73%**

Training Dataset

Tool		miR-abela	
Species	**Total miRNA**		
Human (*homo sapiens*)	140	119	85.00%
Mouse (*mus musculus*)	112	86	76.79%
Rat (*rattus norvegicus*)	72	60	83.33%
Chicken (*gallus gallus*)	54	48	88.89%
Rhesus Macaque (*macaca mullata*)	96	77	80.21%
Dog (*canis familiaris*)	64	52	81.25%
Chimp (*pan troglodytes*)	29	26	89.66%
Cow (*bos taurus*)	46	36	78.26%
Opossum (*monodelphis domestica*)	29	18	62.07%
Zebra Fish (*danio rerio*)	98	75	76.53%
	740	**597**	**80.68%**

Training Dataset

Tool		PROMIR II	
Species	**Total miRNA**		
Human (*homo sapiens*)	140	104	74.29%
Mouse (*mus musculus*)	112	81	72.32%
Rat (*rattus norvegicus*)	72	47	65.28%
Chicken (*gallus gallus*)	54	42	77.78%
Rhesus Macaque (*macaca mullata*)	96	N/A	0.00%
Dog (*canis familiaris*)	64	N/A	0.00%
Chimp (*pan troglodytes*)	29	N/A	0.00%
Cow (*bos taurus*)	46	N/A	0.00%
Opossum (*monodelphis domestica*)	29	N/A	0.00%
Zebra Fish (*danio rerio*)	98	N/A	0.00%
	740(378)	**274**	**72.49%**

Table 5 Results for miRNAMatcher, miR-abela and PROMIR II for the combined datasets

Combined Datasets			
Tool		**miRNAMatcher**	
Species	**Total miRNA**		
All species	**1534**	**1440**	**93.87%**

Combined Datasets			
Tool		**miR-abela**	
Species	**Total miRNA**		
All species	**1534**	**1229**	**80.12%**

Combined Datasets			
Tool		**PROMIR II**	
Species	**Total miRNA**		
All species	**1534(754)**	**525**	**69.63%**

5.2.Tests on genome wide miRNA identification

In order to determine if the algorithm could be applied to large genome scale scanning and to determine how well the algorithm would discard random non-miRNA segments a second test was derived. The dog (canis familiaris) genome was chosen as the target to scan, there are many fewer known miRNA for this species than for humans making it a good choice in terms of the possibility of finding miRNA candidates for known miRNA families. Each chromosome was broken up in to approximately 10MB genomic DNA sections resulting in between seven to eleven data files per chromosome. The data files were then each searched using miRNAMatcher, with all data files belonging to a particular chromosome processed in parallel on a multi-processor machine. Results for three chromosomes (one, two and eight) will be presented here, chromosomes one and eight were chosen as they have a relatively high number of known miRNA and chromosome two as it has very few known miRNA. In terms of performance it is believed that the results are acceptable for genome wide miRNA identification.

Table 6 Chromosome scanning performance results

Chromosome 1	
No. of data files	12
Total Running Time	35.5 days
Parallel Running Time	74 hours
Total no. of bases	125,616,256 per strand
Bases scanned per hour	294183.26 bases per hour
Chromosome 2	
No. of data files	9
Total Running Time	24.5 days
Parallel Running Time	74 hours
Total no. of bases	88,410,189 per strand
Bases scanned per hour	289774.46 bases per hour
Chromosome 8	
No. of data files	8
Total Running Time	22.1 days
Parallel Running Time	74 hours
Total no. of bases	77,315,194 per strand
Bases scanned per hour	290877.32 bases per hour

5.2.1. Canis familiaris Chromosome 1 miRNA identification results

Chromosome 1 has thirteen known miRNA in version 11 of the Sanger miRBase registry (April 2008) (Griffiths-Jones, Grocock, van Dongen, Bateman, & Enright, 2006), (Griffiths-Jones, Saini, van Dongen, & Enright, 2008), three of these fall in to the category of not having enough miRNA family members in the training data to confidently build a classifier. The human chromosome 1 has forty seven known miRNA.

Table 7 Chromosome 1 raw scan results showing no. of candidate miRNA identified

miRNA Family	No. Candidate miRNA Identified	Classifier Quality (unseen dataset)
MIPF0000095	1	100% - 100% - 100%
MIPF0000017	1	100% - 100% - 99%
MIPF0000036	1	99% - 99% - 99%
MIPF0000041	1	100% - 100% - 100%
MIPF0000042	1	99% - 99% - 99%
MIPF0000046	1	99% - 88% - 88%
MIPF0000027	1	100% - 100% - 100%
MIPF0000025	1	100% - 100% - 100%
MIPF0000022	1	100% - 100% - 100%
MIPF0000038	2	100% - 99% - 99%
MIPF0000114	3	100% - 100% - 100%
MIPF0000028	4	100% - 100% - 100%
MIPF0000031	11	100% - 100% - 100%
MIPF0000002	12	99% - 99% -98%
MIPF0000317	15	92% - 88% - 81%
MIPF0000019	15	95% - 95% - 95%
MIPF0000006	23	99% - 97% - 97%
MIPF0000013	55	100% - 99% - 99%
MIPF0000014	60	93% - 93% - 93%
MIPF0000113	227	99% - 99% - 95%
MIPF0000130	739	80% - 74% - 74%
MIPF0000001	1854	98% - 97% - 96%
MIPF0000316	3017	96% - 96% - 94%
MIPF0000018	3085	78% - 66% - 60%

It was immediately noted that the last eight families in Table 7 seemed to be over represented, the regular expressions for these families were investigated and it was found that these families suffered from a flaw in the genetic algorithm training system where

families with predominantly overlapping motifs were resulting in one of the regular expressions to contain a motif pattern that consists of continuous characters such as **AGCUGGAU**, this pattern of around seven bases is too easily found randomly. It is a coincidence in these families that after applying our matching rule, where the number of patterns less three is required to indicate a match, that only this single pattern is being required. The amount of negative bias for a continuous motif needs to be increased to prevent this problem and further solutions are discussed in later sections. The regular expressions for these families were modified to remove the continuous motifs and the results post processed, the results are shown below.

Table 8 Chromosome 1 post processed scan results showing no. of candidate miRNA identified

miRNA Family	No. Candidate miRNA Identified	Classifier Quality (unseen dataset)
MIPF0000095	1	100% - 100% - 100%
MIPF0000017	1	100% - 100% - 99%
MIPF0000036	1	99% - 99% - 99%
MIPF0000041	1	100% - 100% - 100%
MIPF0000042	1	99% - 99% - 99%
MIPF0000046	1	99% - 88% - 88%
MIPF0000027	1	100% - 100% - 100%
MIPF0000025	1	100% - 100% - 100%
MIPF0000022	1	100% - 100% - 100%
MIPF0000038	2	100% - 99% - 99%
MIPF0000114	3	100% - 100% - 100%
MIPF0000028	4	100% - 100% - 100%
MIPF0000031	11	100% - 100% - 100%
MIPF0000002	12	99% - 99% -98%
MIPF0000317	15	92% - 88% - 81%
MIPF0000019	15	95% - 95% - 95%
MIPF0000006	10	99% - 97% - 97%
MIPF0000013	9	100% - 99% - 99%
MIPF0000014	26	93% - 93% - 93%
MIPF0000113	119	99% - 99% - 95%
MIPF0000130	348	80% - 74% - 74%
MIPF0000001	758	98% - 97% - 96%
MIPF0000316	944	96% - 96% - 94%
MIPF0000018	272	78% - 66% - 60%

There is a significant improvement after post processing the last eight miRNA families, retraining of the genetic algorithm however would be the best approach to ensure that the manual regular expression modifications do not result in other problems such as classifiers matching miRNA from other miRNA families.

All eleven of the known miRNA falling into the families for which classifiers have been built have been identified by our scan of the chromosome data. Additionally one miRNA was found for a miRNA family that had less than 8 members, which was originally discarded in data pre-processing, see last paragraph of section 4.1.1 for reasoning.

5.2.1.1. Results for matched known miRNA on Chromosome 1

Figure 6: Chromosome 1 - Matched miRNA cfa-mir-122

ID	Position	Known Structure
cfa-mir-122 Family: MIPF0000095	20601556-20601613 [-]	--ugg c --u c aguguga aaugguguuug gu c ||||||| ||||||||||| || ucacacu uuaccgcaaac ca a auaaa a uau a
Motif Position: 20601613, Motif Family: MIPF0000095		
Discovered Structure		

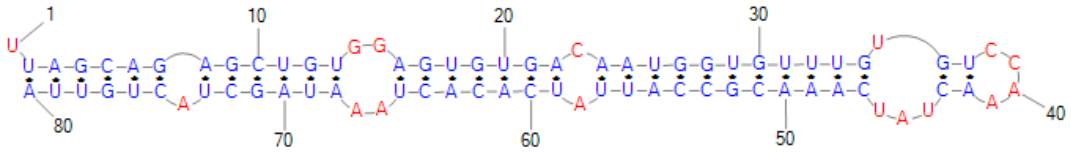


Figure 7: Chromosome 1 - Matched miRNA cfa-mir-24-1

ID	Position	Known Structure
cfa-mir-24-1 Family: MIPF0000041	74734229-74734290 [-]	--g g a ua ucuca gu ccu cugagcuga ucagu u || ||| ||||||||| ||||| u ca gga gacuugacu gguca g gga a c -c cacau
Motif Position: 74734256, Motif Family: MIPF0000041		
Discovered Structure		

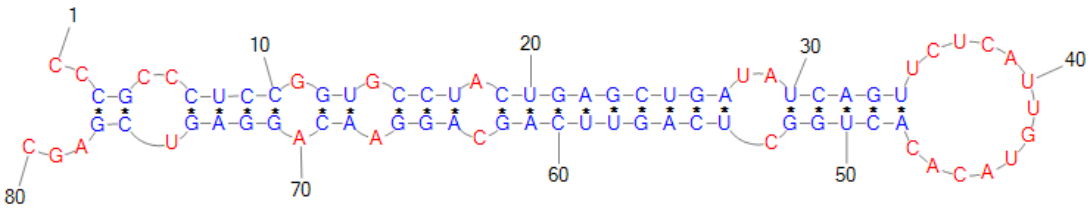


Figure 8: Chromosome 1 - Matched miRNA cfa-mir-27b

ID	Position	Known Structure
cfa-mir-27b Family: MIPF0000036	74734738-74734800 [-]	-- auug ugac agagcuuagcug gugaacag uggu |||||||||||| |||||||| ||| u ucuugaaucggu cacuuguu gccu cg --ga --uc
Motif Position: 74734752, Motif Family: MIPF0000036		
Discovered Structure		

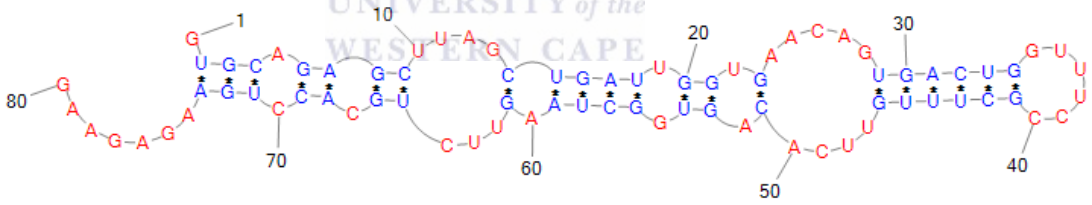


Figure 9: Chromosome 1 - Matched miRNA cfa-mir-23b

ID	Position	Known Structure
cfa-mir-23b Family: MIPF0000027	74734974-74735027 [-]	-- - c gugacu guuccuggca ug ugauuu u |||||||||| || |||||| uagggaccgu ac acuaaa a au u - auuaga
Motif Position: 74734996, Motif Family: MIPF0000027		
Discovered Structure		

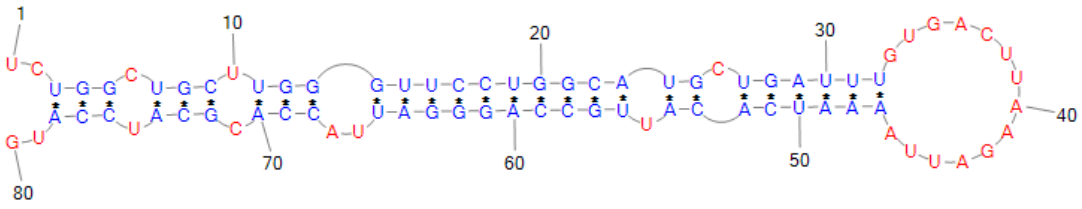


Figure 10: Chromosome 1 - Matched miRNA cfa-mir-7-1

ID	Position	Known Structure
cfa-mir-7-1 Family: MIPF0000022	78565664-78565727 [+]	-- a a u -- a ugg agacu gugauuu guugu uuuag u ||| ||||| |||||| |||||| ||||| acc ucuga cacuaaa caacag aaauc a au g - - cu a
Motif Position: 78565690, Motif Family: MIPF0000022		
Discovered Structure		

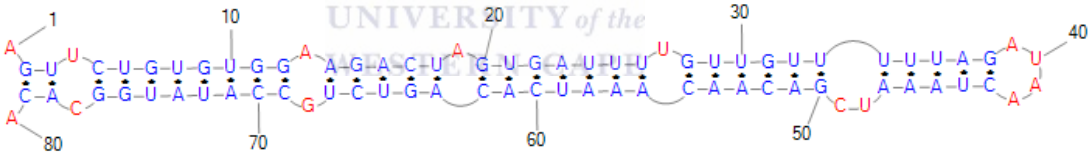


Figure 11: Chromosome 1 - Matched miRNA cfa-mir-204

ID	Position	Known Structure
cfa-mir-204 Family: MIPF0000042	89887249-89887308 [+]	--u a u gagaau ucccuuuguc uccua gccu a |||||||||| ||||| |||| u agggaaacgg agggu cgga a ugc a - ggaagu
Motif Position: 89887273, Motif Family: MIPF0000042		
Discovered Structure		

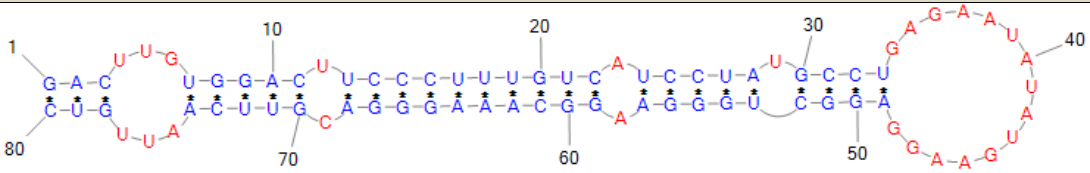


Figure 12: Chromosome 1 - Matched miRNA cfa-mir-101-2

ID	Position	Known Structure
cfa-mir-101-2 Family: MIPF0000046	96260667-96260721 [+]	-- c a guaua gguuaucaugguac g ugcu u ||||||||||||| |||| ucaauagugucaug c augg c ag a - aaagu
Motif Position: 96260721, Motif Family: MIPF0000046		
Discovered Structure		

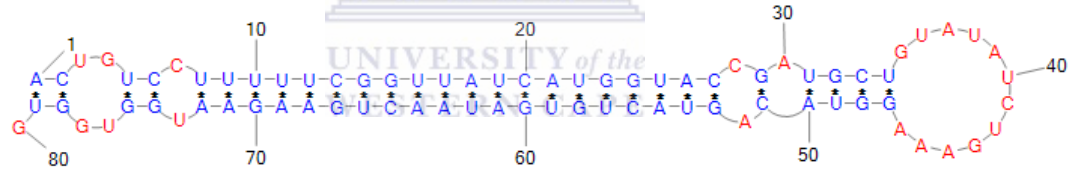


Figure 13: Chromosome 1 - Matched miRNA cfa-let-7f

ID	Position	Known Structure
cfa-let-7f Family: MIPF0000002	100972836- 100972913 [-]	--u ---------- u gagguaguagauuguauaguugu ggggguag g |||||||||||||||||||||||| ||||||| a uuccguuaucuaacauaucaaua ucccauu u ccc gaggacuug u
Motif Position: 100972911, Motif Family: MIPF0000002		
Discovered Structure		

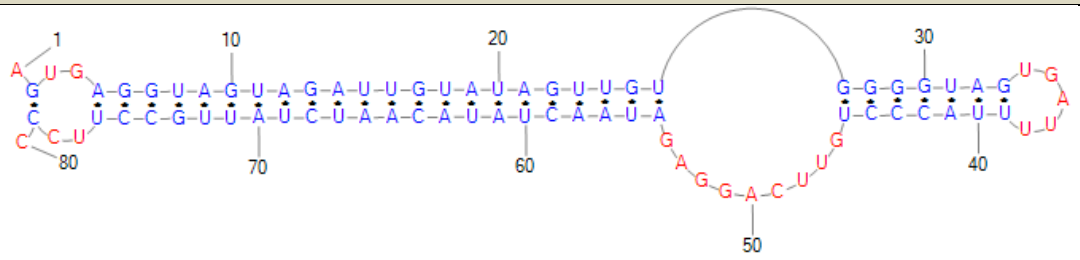


Figure 14: Chromosome 1 - Matched miRNA cfa-mir-125a

ID	Position	Known Structure
cfa-mir-125a Family: MIPF0000017	108280390-108280449 [-]	--uc ug c ua ---- a cc aga ccuu accuguga gg c || ||| |||| |||||||| || gg ucu ggag uggacacu cc a ccga gu u -- ggga u
Motif Position: 108280450, Motif Family - MIPF0000017		
Discovered Structure		

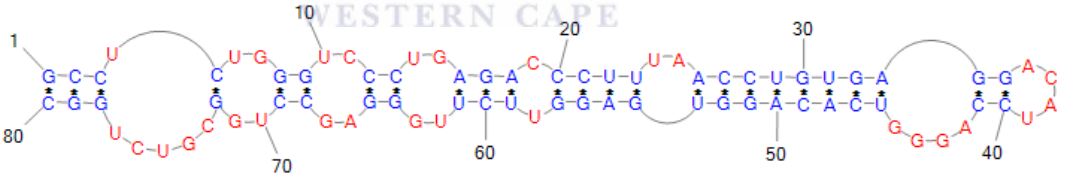


Figure 15: Chromosome 1 - Matched miRNA cfa-let-7e

ID	Position	Known Structure
cfa-let-7e Family: MIPF0000002	108280857- 108280923 [-]	--u g u ----g a a gag uaggagguuguauagu ga g gg c ||| |||||||||||||| || | || uuc auccuccggcauauca cu c cc a ccu g - agagg a c
Motif Position: 108280925, Motif Family - MIPF0000002		
Discovered Structure		

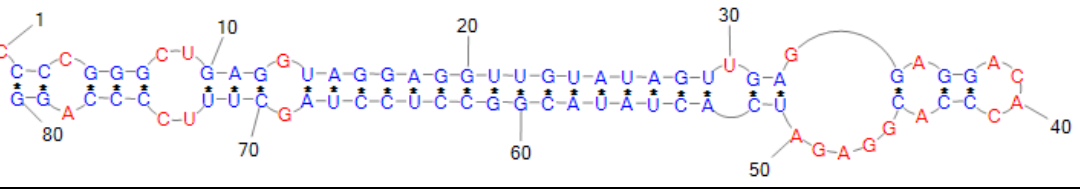
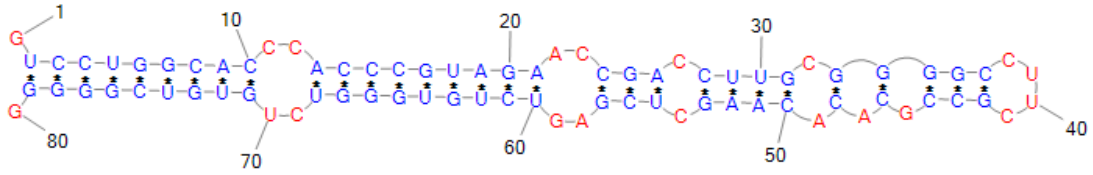


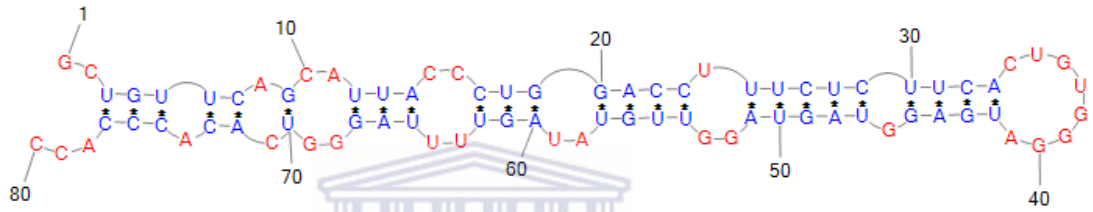
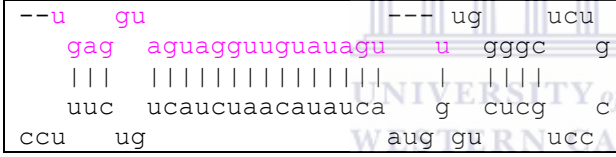
Figure 16: Chromosome 1 - Matched miRNA cfa-mir-99b

ID	Position	Known Structure
cfa-mir-99b Family: MIPF0000025	108281042- 108281101 [-]	--c ac c c - - c acccguaga cga cuug g g ggc u ||||||||| ||| |||| | | ||| uggggucu gcu gaac c c ccg u guc ga c a a g c
Motif Position: 108281109, Motif Family - MIPF0000025		
Discovered Structure		
		

5.2.1.2. Results for matched unknown miRNA on Chromosome 1

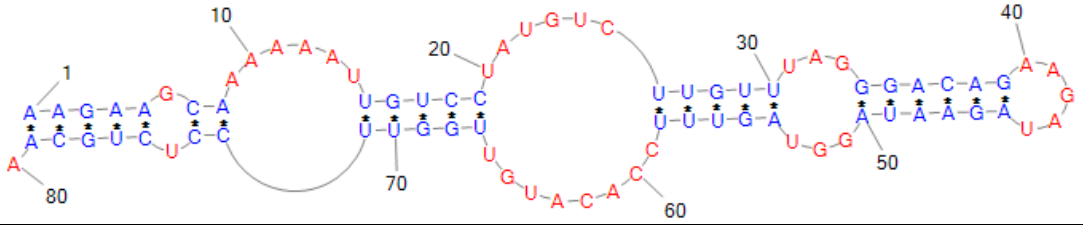
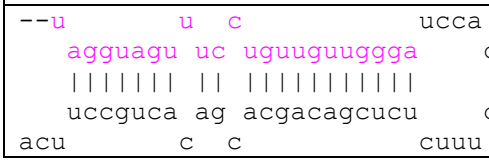
The predicted miRNA candidate below (Figure 17) has a strong relationship to the **let-7a** miRNA, it has a perfect match for the mature miRNA and the prediction is for the same miRNA family. The noticeable differences are that the mature miRNA section is on the 3' rather than the 5' end of the miRNA pre-cursor. The known **let-7a** miRNA for *canis-familiaris* is on chromosome 14 while this candidate appears on chromosome 1.

Figure 17: Chromosome 1 - Unmatched miRNA related to cfa-let-7a

Motif Position: 100973290, Motif Family - MIPF0000002
Discovered Structure

Related Known miRNAs: cfa-let-7a, Family MIPF0000002
Position: Chromosome 10: 23089107-23089175 [-]


The predicted miRNA candidate below (Figure 18) has a reasonably strong relationship to the **miR-196b** miRNA, it has a close match for the mature miRNA (differing by 3 bases) and the prediction is for the same miRNA family. The noticeable differences are that the mature miRNA section is on the 3' rather than the 5' end of the miRNA pre-cursor. The known **miR-196b** miRNA for *canis-familiaris* is on chromosome 14 while this candidate appears on chromosome 1.

Figure 18: Chromosome 1 - Unmatched miRNA related to cfa-mir-196b

Motif Position: 42989910, Motif Family - MIPF0000031
Discovered Structure

Related Known miRNAs: cfa-miR-196b, Family MIPF0000031
Position: Chromosome 14: 43299510-43299567 [-]


Some examples below of miRNA candidates that do not have any relationship to known miRNA.

Figure 19: Chromosome 1 - Unmatched miRNA with no relationship to known miRNA

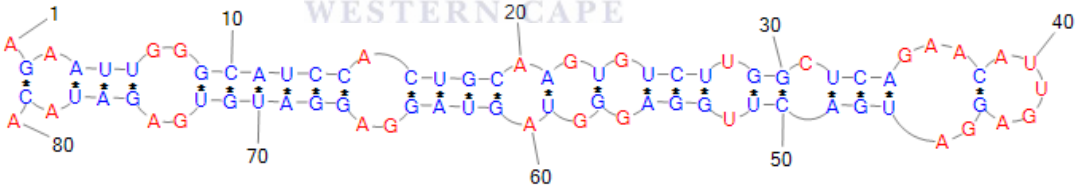
Motif Position: 83165231, Motif Family - MIPF0000028
Discovered Structure

Related Known miRNAs
None

Figure 20: Chromosome 1 - Unmatched miRNA with no relationship to known miRNA

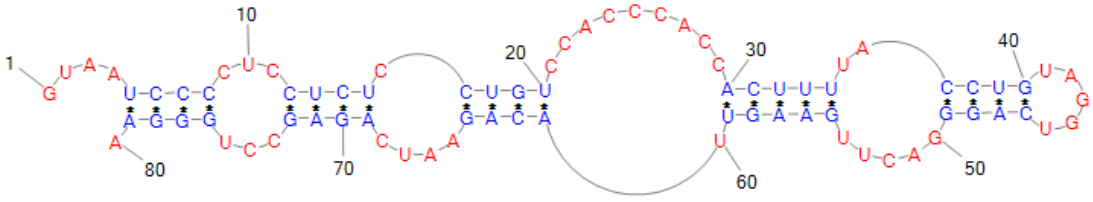
Motif Position: 34570412, Motif Family - MIPF0000317
Discovered Structure

Related Known miRNAs
None

Figure 21: Chromosome 1 - Unmatched miRNA with no relationship to known miRNA

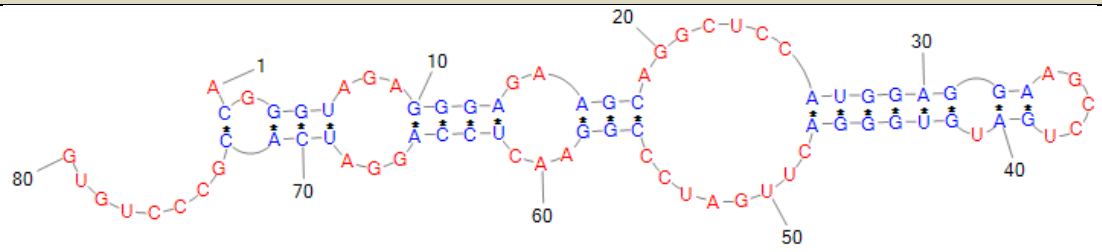
Motif Position: 89320987, Motif Family - MIPF0000317
Discovered Structure

Related Known miRNAs
None

Figure 22: Chromosome 1 - Unmatched miRNA with no relationship to known miRNA

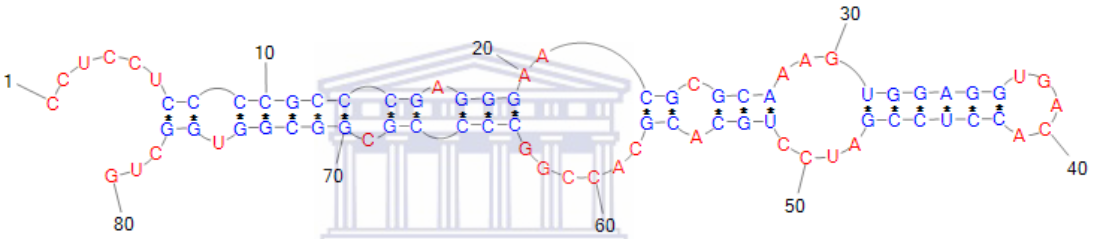
Motif Position: 93776431, Motif Family - MIPF0000114
Discovered Structure

Related Known miRNAs
None

Figure 23: Chromosome 1 - Unmatched miRNA with no relationship to known miRNA

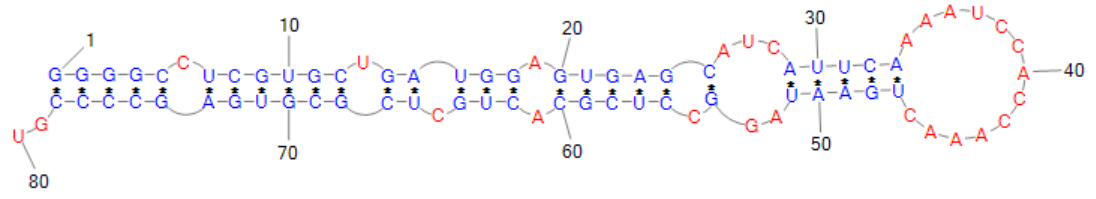
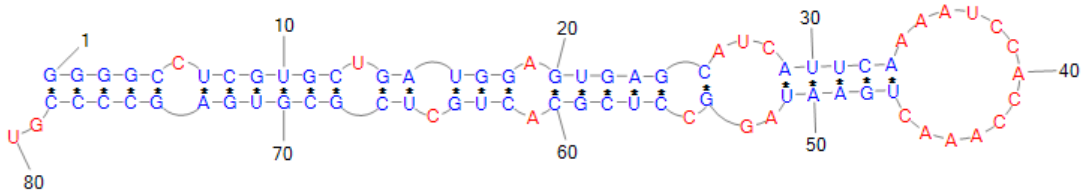
Motif Position: 71447764, Motif Family - MIPF0000114
Discovered Structure

Related Known miRNAs
None

Figure 24: Chromosome 1 - Unmatched miRNA with no relationship to known miRNA

Motif Position: 71447764, Motif Family - MIPF0000114
Discovered Structure

Related Known miRNAs
None

5.2.2. Canis familiaris Chromosome 2 miRNA identification results

Chromosome 2 has two known miRNA in version 11 of the Sanger miRBase registry (April 2008) (Griffiths-Jones, Grocock, van Dongen, Bateman, & Enright, 2006), (Griffiths-Jones, Saini, van Dongen, & Enright, 2008), one of these falls in to the category of not having enough miRNA family members in the training data to confidently build a classifier. The human chromosome 2 has twenty four known miRNA.

Table 9 Chromosome 2 raw scan results showing no. of candidate miRNA identified

miRNA Family	No. Candidate miRNA Identified	Classifier Quality (unseen dataset)
MIPF0000034	1	100% - 100% - 100%
MIPF0000022	1	100% - 100% - 100%
MIPF0000038	1	100% - 100% - 99%
MIPF0000075	1	100% - 100% - 99%
MIPF0000039	2	96% - 93% - 93%
MIPF0000028	3	100% - 100% - 100%
MIPF0000002	3	99% - 98% - 96%
MIPF0000317	6	92% - 88% -81%
MIPF0000031	8	100% - 100% - 100%
MIPF0000006	8	99% - 97% - 97%
MIPF0000019	13	95% - 95% - 95%
MIPF0000014	37	93% - 93% - 93%
MIPF0000013	57	100% - 99% - 99%
MIPF0000113	146	99% - 99% - 95%
MIPF0000130	415	80% - 74% -74%
MIPF0000001	1287	98% - 97% - 96%
MIPF0000316	1763	96% - 96% - 94%
MIPF0000018	2132	78% - 66% - 60%

Table 10 Chromosome 2 post processed scan results showing no. of candidate miRNA identified

miRNA Family	No. Candidate miRNA Identified	Classifier Quality (unseen dataset)
MIPF0000034	1	100% - 100% - 100%
MIPF0000022	1	100% - 100% - 100%
MIPF0000038	1	100% - 100% - 99%
MIPF0000075	1	100% - 100% - 99%
MIPF0000039	2	96% - 93% - 93%
MIPF0000028	3	100% - 100% - 100%
MIPF0000002	3	99% - 98% - 96%
MIPF0000317	6	92% - 88% -81%
MIPF0000031	8	100% - 100% - 100%
MIPF0000006	8	99% - 97% - 97%

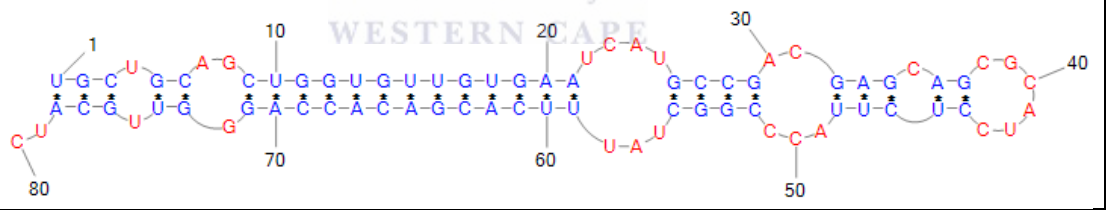
MIPF0000019	6	95% - 95% - 95%
MIPF0000014	17	93% - 93% - 93%
MIPF0000013	9	100% - 99% - 99%
MIPF0000113	72	99% - 99% - 95%
MIPF0000130	224	80% - 74% -74%
MIPF0000001	505	98% - 97% - 96%
MIPF0000316	553	96% - 96% - 94%
MIPF0000018	157	78% - 66% - 60%

Once again there is a significant improvement after post processing the last eight miRNA families.

5.2.2.1. Results for matched known miRNA on Chromosome 2

Figure 25: Chromosome 2 - Matched miRNA cfa-mir-138b

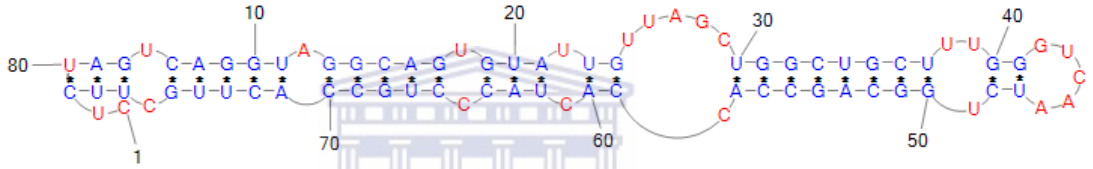
ID	Position	Known Structure
cfa-mir-138b Family: MIPF0000075	62334160- 62334228 [-]	--ag ucau -ac c cg cugguguugugaa gccg gag ag c |||||||||||||| |||| ||| || gaccacagcacuu cggc uuc uc a uugg -uau cca - cu
Motif Position: 62334218, Motif Family: MIPF0000075		
Discovered Structure		



5.2.2.2. Results for unknown miRNA on Chromosome 2

The miRNA candidate below (Figure 26) is a very close match to the mir-449b family of miRNA, for which no known miRNA exists for *canis-familiaris*. The candidate miRNA is approximately 100 bases downstream from the known cfa-mir-449 miRNA, which is similar to other species for which both mir-449 and mir-449b are known. The secondary structure for the human hsa-mir-449b is shown below and is a very good match for the predicted structure of our candidate.

Figure 26: Chromosome 2 - Unmatched miRNA related to hsa-mir-449b

Motif Position: 45390562, Motif Family - MIPF0000039
Discovered Structure

Related Known miRNAs: hsa-mir-449b, Family MIPF0000133
Position: Chromosome 2: ~100-200 bases > 45390354-45390444 [-]
---------uga u u a u uu a c u gu cc gaa caggu ggcag gua guu g uggcugcu gg c || ||| ||||| ||||| ||| ||| | ||||||| || gg cuu guuca ccguc cau caa c accgacga cu a ucuucuuaaaua u c - c -- - - - ga

Results for candidates that have no relation to known miRNA will not be shown here as the results for chromosome 1 illustrate the results sufficiently.

5.2.3. Canis familiaris Chromosome 8 miRNA identification results

Chromosome 8 has twenty two known miRNA in version 11 of the Sanger miRBase registry (April 2008) (Griffiths-Jones, Grocock, van Dongen, Bateman, & Enright, 2006), (Griffiths-Jones, Saini, van Dongen, & Enright, 2008), eight of these fall in to the category of not having enough miRNA family members in the training data to correctly build a classifier. The human chromosome 8 has twenty eight known miRNA.

Table 11 Chromosome 8 raw scan results showing no. of candidate miRNA identified

miRNA Family	No. Candidate miRNA Identified	Classifier Quality (unseen dataset)
MIPF0000189	1	99% - 99% - 99%
MIPF0000039	1	96% - 93% - 93%
MIPF0000110	1	99% - 99% - 99%
MIPF0000178	2	100% - 100% - 99%
MIPF0000038	3	100% - 100% - 99%
MIPF0000028	3	100% - 100% - 100%
MIPF0000126	4	99% - 96% - 96%
MIPF0000091	4	100% - 100% - 100%
MIPF0000317	5	92% - 88% - 81%
MIPF0000002	5	99% - 98% -98%
MIPF0000031	7	100% - 100% - 100%
MIPF0000019	12	95% - 95% - 95%
MIPF0000006	20	98% - 97% - 97%
MIPF0000014	35	93% - 93% - 93%
MIPF0000013	39	100% - 99% - 99%
MIPF0000113	107	99% - 99% - 95%
MIPF0000130	456	80% - 74% - 74%
MIPF0000001	1181	98% - 97% - 96%
MIPF0000316	1731	96% - 96% - 94%
MIPF0000018	2008	78% - 66%- 60%

Table 12 Chromosome 8 post processed scan results showing no. of candidate miRNA identified

miRNA Family	No. Candidate miRNA Identified	Classifier Quality (unseen dataset)
MIPF0000189	1	99% - 99% - 99%
MIPF0000039	1	96% - 93% - 93%
MIPF0000110	1	99% - 99% - 99%
MIPF0000178	2	100% - 100% - 99%
MIPF0000038	3	100% - 100% - 99%
MIPF0000028	3	100% - 100% - 100%
MIPF0000126	4	99% - 96% - 96%
MIPF0000091	4	100% - 100% - 100%

MIPF0000317	5	92% - 88% - 81%
MIPF0000002	5	99% - 98% -98%
MIPF0000031	7	100% - 100% - 100%
MIPF0000019	12	95% - 95% - 95%
MIPF0000006	7	98% - 97% - 97%
MIPF0000014	16	93% - 93% - 93%
MIPF0000013	6	100% - 99% - 99%
MIPF0000113	43	99% - 99% - 95%
MIPF0000130	233	80% - 74% - 74%
MIPF0000001	507	98% - 97% - 96%
MIPF0000316	112	96% - 96% - 94%
MIPF0000018	236	78% - 66%- 60%

Once again there is a significant improvement after post processing the last eight miRNA families.

5.2.3.1. Results for matched known miRNA on Chromosome 8

Figure 27: Chromosome 8 - Matched miRNA cfa-mir-345

ID	Position	Known Structure
cfa-mir-345 Family: MIPF0000189	71630689-71630748 [+]	--ugcu c u ug u gacucuuagu cag gcucg a g cuggggauca guc cgggu u g ggaggu a c gg c
Motif Position: 71630744, Motif Family: MIPF0000189		
Discovered Structure		

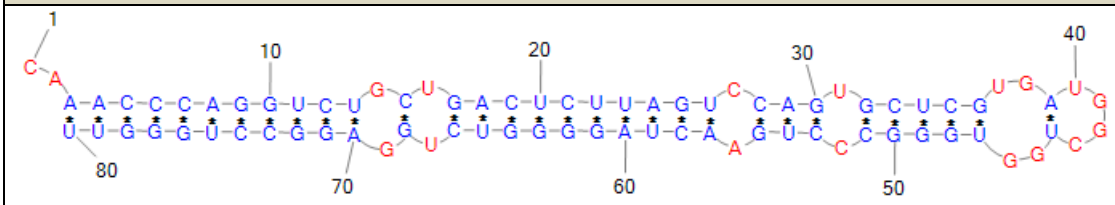


Figure 28: Chromosome 8 - Matched miRNA cfa-mir-379

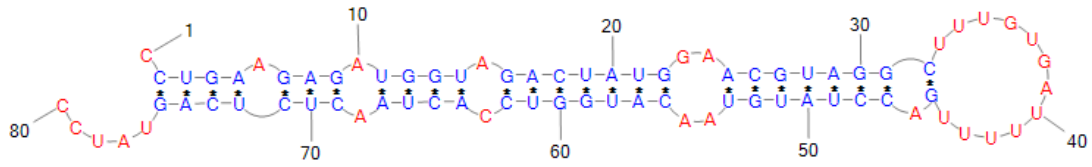
ID	Position	Known Structure
cfa-mir-379 Family: MIPF0000189	72296541-72296599 [+]	-- a ga - uuugu uggu gacuaug acguagg c g |||| ||||||| ||||||| | auca cugguac uguaucc g a ca c aa a uuuuu
Motif Position: 72296596, Motif Family: MIPF0000189		
Discovered Structure		
		

Figure 29: Chromosome 8 - Matched miRNA cfa-mir-411

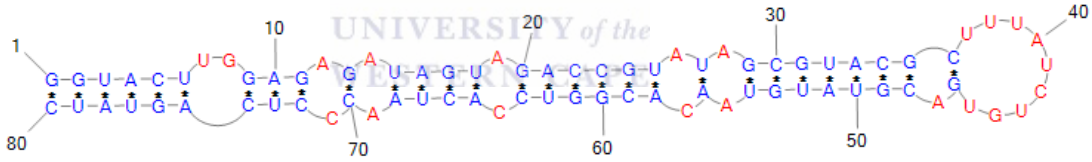
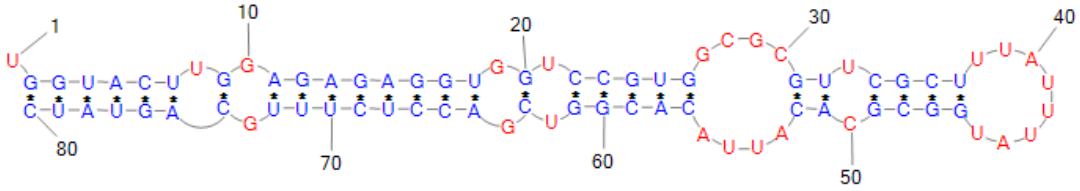
ID	Position	Known Structure
cfa-mir-411 Family: MIPF0000126	72297827-72297884 [+]	--a a a a - uuu uagu gaccgu u gcguacg c a |||| |||||| | ||||||| | u auca cuggca a uguaugc g c cca c c a a ugu
Motif Position: 72297863, Motif Family: MIPF0000018		
Discovered Structure		
		

Figure 30: Chromosome 8 - Matched miRNA cfa-mir-323

ID	Position	Known Structure
cfa-mir-323 Family: MIPF0000018	72300227-72300282 [+]	-- g u gcgc u uua aggu g ccgug gu cgcu u |||| | ||||| || |||| ucca c ggcac ca gcgg u uc g u auua c uau
Motif Position: 72300300, Motif Family: MIPF0000018		
Discovered Structure		



The matched miRNA below (Figure 31) has the incorrect family but has an exact sequence match, however the secondary structure prediction is not accurate as the hairpin loop occurs too far down the sequence segment.

Figure 31: Chromosome 8 – Partially Matched miRNA cfa-mir-329

ID	Position	Known Structure
cfa-mir-329 Family: MIPF0000110	72301588-72301647 [+]	-- uu uuc u a agagguu cugggu uguuuc uuc u ||||||| |||||| |||||| ||| ucuccaa gaccca acaaag agg g uu uu --c u a
Motif Position: 72301662, Motif Family: MIPF0000018		
Discovered Structure		

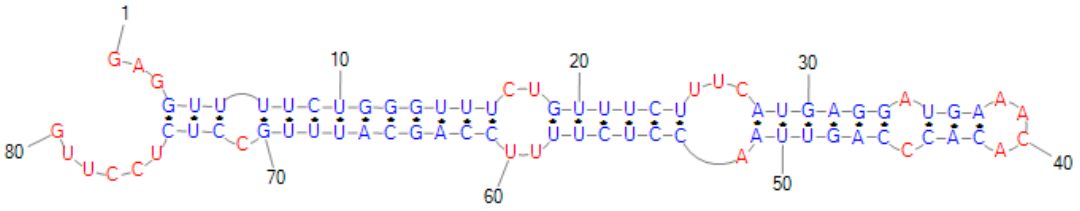


Figure 32: Chromosome 8 - Matched miRNA cfa-mir-543

ID	Position	Known Structure
cfa-mir-543 Family: MIPF0000110	72306139-72306196 [+]	-- u - ---u u uu gaagu gc ccgug uuuu ucgc u ||||| || ||||| |||| |||| a cuuca cg ggcgc aaag agug u uu - u uuac c u
Motif Position: 72306203, Motif Family: MIPF0000110		
Discovered Structure		

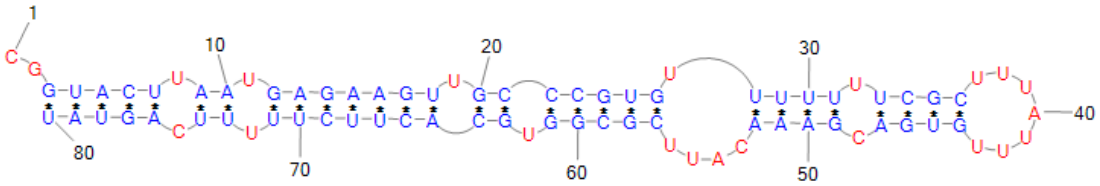


Figure 33: Chromosome 8 - Matched miRNA cfa-mir-376-2

ID	Position	Known Structure
cfa-mir-376-2 Family: MIPF0000091	72313232-72313289 [+]	-- a u -- g gu gauuuuccu cuaugauua cgu u || |||||||| |||||||| ||| u ca cuaaaagga gauacuaau gua u ug c - ug g
Motif Position: 72313306, Motif Family: MIPF0000091		
Discovered Structure		

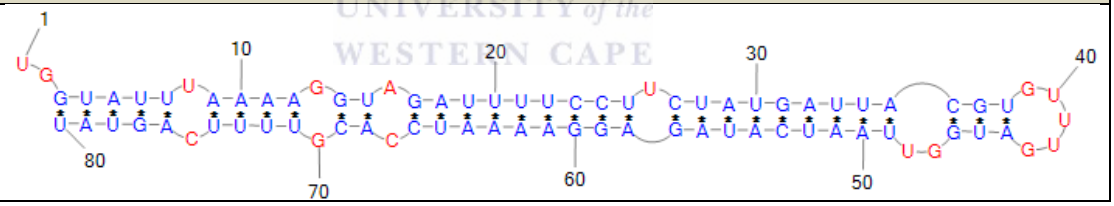


Figure 34: Chromosome 8 - Matched miRNA cfa-mir-487

ID	Position	Known Structure
cfa-mir-487 Family: MIPF0000018	72318229-72318286 [+]	-- uuau c u uuuua gugg cccugu c guucg c |||| |||||| | ||||| cacc gggaca g uaagc u uu uacu u c uguac
Motif Position: 72318228, Motif Family: MIPF0000018		
Discovered Structure		

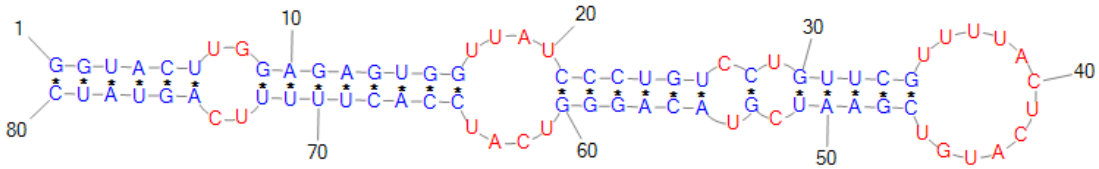
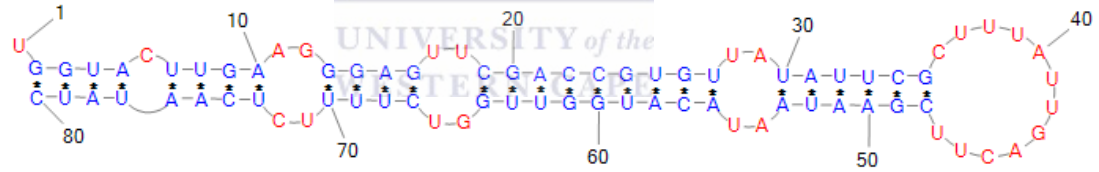


Figure 35: Chromosome 8 - Matched miRNA cfa-mir-369

ID	Position	Known Structure
cfa-mir-369 Family: MIPF0000018	72336806-72336861 [+]	-- uu ua cuuua ag cgaccgugu uauucg u || |||||||| |||||| uc guugguaca auaagc u uu ug ua uucag
Motif Position: 72336879, Motif Family: MIPF0000018		
Discovered Structure		



5.2.3.2. Results for unknown miRNA on Chromosome 8

Figure 36: Chromosome 8 - Unmatched miRNA related to hsa-mir-208a

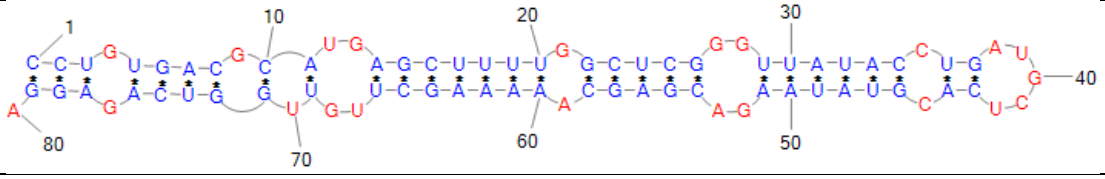
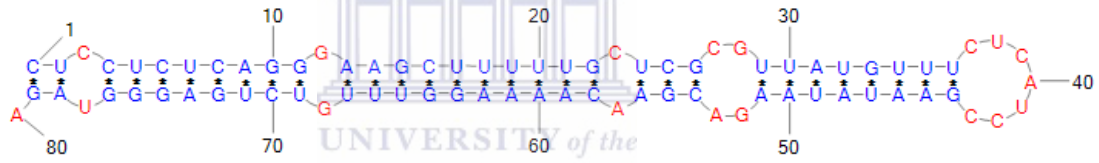
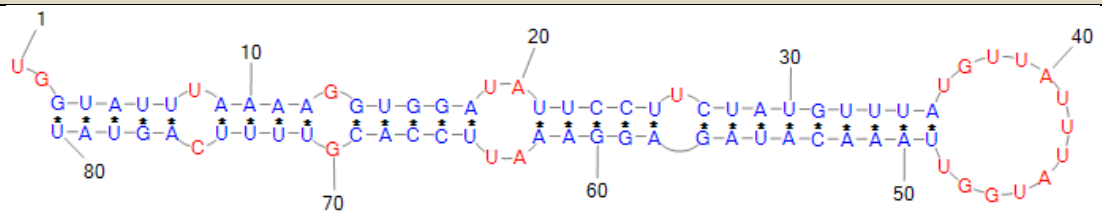
Motif Position: 6640095, Motif Family - MIPF0000178
Discovered Structure

Related Known miRNAs: hsa-mir-208a, Family MIPF0000178
Position: Chromosome 14: 22927645-22927715 [-]
u g g c gg c a gac ggcgagcuuuu gc cg uuauac ug u ||| |||||||||| || || ||||| || g cug uuguucgaaaa cg gc aauaug ac c a g a a ag c u

Figure 37: Chromosome 8 - Unmatched miRNA related to hsa-mir-208b

Motif Position: 6668353, Motif Family - MIPF0000178
Discovered Structure

Related Known miRNAs: hsa-mir-208b, Family MIPF0000178
Position: Chromosome 14: 22957036-22957112 [-]
--c g c aa cu cucucagg aagcuuuuug ucg uuauguuu g |||||||| ||||||||| ||| ||||||| a gggagucu uuuggaaaac agc aauauaag u gac g a ag cc

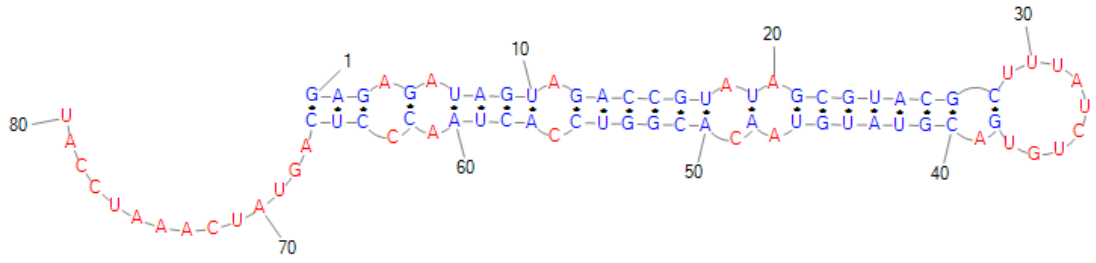
The miRNA candidate below (Figure 38) has a genomic location that falls between the known miRNA cfa-mir-376-3 and cfa-mir-376-2 and has an exact mature miRNA match with hsa-mir-376c as well as a perfect match for the stem loop section of the pre-cursor miRNA.

Figure 38: Chromosome 8 - Unmatched miRNA related to hsa-mir-376c

Motif Position: 72312918, Motif Family - MIPF0000091
Discovered Structure

Related Known miRNAs: hsa-mir-376c, Family MIPF0000091
Position: Chromosome 14: 100575780-100575845 [+]
g ua u uguua aaaa gugga uuccu cuauguuua u |||| ||||| ||||| ||||||||| u uuuu caccu aagga gauacaaau u g ua – uggua

The miRNA candidate below (Figure 39) has a strong similarity to the mature miRNA hsa-mir-379, both the candidate and the known miRNA have the mature miRNA located on the 3' arm of the pre-cursor miRNA and both are in close proximity to their corresponding species specific versions of mir-411 and mir-329.

Figure 39: Chromosome 8 - Unmatched miRNA related to hsa-mir-379

Motif Position: 72297966, Motif Family - MIPF0000126
Discovered Structure

Related Known miRNAs: hsa-mir-379, Family MIPF0000126
Position: Chromosome 14: 100558156-100558222 [+]
a a ga – uu u agag uggu gacuaug acguagg cg a g |||| |||| ||||||| ||||||| || | ucuc auca cugguac uguaucc gu u a a c aa a cu u

Results for candidates that have no relation to known miRNA will not be shown here as the results for chromosome 1 illustrate the results sufficiently.

6. Discussion

The initial results showed that classifiers for some miRNA families were identifying too many miRNA candidates, for example the miRNA family MIPF0000018 classifiers (Table 7) identified 3085 candidates on chromosome one and seeing as though this is similar in size to about half the known miRNA it seems reasonably unlikely. The problem was soon attributed to classifier regular expressions built from overlapping motifs being promoted as the single regular expression used as the algorithm only requires the number of regular expressions – 3 to be matched. In some instances only four regular expressions were being used one of which was built from overlapping motifs. This regular expression matching a short continuous sequence of RNA too easily matches random sequence data. As a simple fix the offending regular expressions were removed and the results post processed with the new classifiers, the miRNA family MIPF0000018 classifiers (Table 8) only identified 272 of the original 3085 as miRNA candidates, a significant improvement. The improvement will be much better if the classifiers are recreated by retraining the genetic algorithm with a rule to either completely exclude overlapping motifs or to require them to exist in combination with separated motifs. The rule of the number of regular expressions – 3 being required for a miRNA candidate match clearly cannot be universally applied to all classifiers and solutions are suggested in the next section.

For chromosomes one and two all known miRNA were matched for miRNA families where sufficient training data existed to build classifiers, i.e. more than 8 members per family were available. For chromosome eight, more than fifty percent were matched. The poorer performance of chromosome eight can be attributed predominantly to the fact that a quarter of the known miRNA on chromosome eight belong to a family for which the worst performing classifier has been built. This particular family of miRNA contains motifs that naturally overlap and have many repeated motifs. The current genetic algorithm was not

able to build good quality classifiers for these cases with most training runs resulting in classifiers only able to recognise sixty percent or less of the training data with a high likelihood of matching random data due to poor motif separation.

It is encouraging that there are a number of results for unknown *canis-familiaris* miRNA that match known miRNA in other species. This is especially evident on chromosome eight where there are very strong candidates and as results are presented for only the miRNA families with under fifteen candidates there is a good likelihood that more homologues will be found when processing all the candidates for matches with known miRNA of other species. It was not felt worthwhile to perform this analysis until several problems identified in this study were rectified; these issues are detailed in the next section.

There were several instances where exact or near exact mature miRNA sequence matches were identified by the classifiers in locations that are different from their known locations. Examples of this can be seen in the results for chromosome one, mature miRNA matches for let7a and mir-196b are found on different chromosomes to their known locations and on the 3' rather than 5' arm of the pre-cursor hairpin. They are both located correctly from the hairpin to be processed by the Drosha and Dicer mechanisms and may represent valid miRNA locations.

The results of scanning several chromosomes show how the simplicity of using three regular expression based scanners achieves excellent linear time performance that can easily be parallelised. Around 250,000,000 bases were scanned in 74 hours with more performance available by using more parallel processes. The highly portable regular expressions can also be easily used from many different programming platforms and so can easily be incorporated by different research groups on their chosen platform.

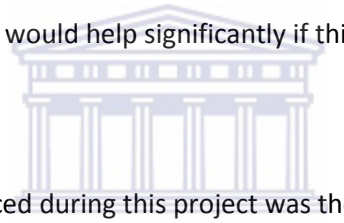
7. Further Problems

Apart from the problems discovered as part of this project such as the problem with overlapping motifs being promoted there are a few others that need further thought in order to improve the usefulness of miRNAMatcher.

Some miRNA family classifiers performed significantly worse than others and these fell into one of two categories, their common motifs either clustered as overlapping features or they contained many repeated motifs. In the first case if a high number of overlapping motifs are chosen to represent a family the result is a classifier that is defined by a number of short continuous RNA segments which can too easily be matched to random RNA. The solution would be to either try downgrade or entirely disallow overlapping motifs or if not possible for a family try build "super" motifs consisting of a number of overlapping motifs separated by some distance just as one would for a regular simple motif. This algorithm would need to be applied as part of the genetic algorithms fitness function. The second problem is more difficult to resolve and would take some further investigation. Currently when the known miRNA precursors for a miRNA family are processed each motif is given an identifying number, and distances are calculated between motifs, however when a motif with the same set of bases is found it is given the same identifying number as before so that a common set of motifs and distances can be found for a particular miRNA family. However if the motif with the same set of bases is actually in a completely different position relative to the other motifs the distance calculations become inaccurate. The solution is to identify when a common motif needs to be given its own identifier. At this stage the only obvious solution would be to post process the related motif distance data to look for motifs that are repeated and then determine if they belong to the current identifier by analysing their position relative to other motifs. This would be a reasonably

slow process to perform, but as it is done during the data pre-processing cycle that may be acceptable.

A small problem was also discovered with finding the exact location of a miRNA pre-cursor after finding a motif match. The motif match is found in a one hundred base sliding window which could theoretically lead to the actual mature miRNA sequence appearing one hundred bases up or down stream from the matched motif. This is predominantly a problem when several miRNA are clustered together where the proposed region could overlap multiple miRNA candidates. A solution other than incrementally expanding the search area outwards is not immediately apparent, however recording knowledge of whether a set of motifs are located towards the beginning, middle or end of the known motifs for an miRNA family would help significantly if this could be easily obtained during the training phase.



The biggest issue experienced during this project was the rule of allowing the classifiers to record a match if three less than the number of available regular expressions represented matches. While this helped some miRNA families enormously the overall effect was quite negative, especially since some classifiers contained overlapping motifs. The best approach is not to apply a global rule to all classifiers but to incorporate a managed decrement of the number of required matches into the genetic algorithm on a per family basis. This would allow families that benefit from not requiring every regular expression to match a chance to be optimised differently to miRNA families that do not require such rules.

8. Conclusion

miRNAMatcher has achieved its goal of providing a fast and efficient way to identify miRNA candidates on a genomic scale. miRNAMatcher successfully avoided the use of computationally intensive secondary structure calculations used in many existing miRNA discovery tools. A few problems were identified with the version produced by this research however the results were quite promising with a high number of known miRNA identified along with several high quality candidates for miRNA homologues in other species. The ability to apply miRNAMatcher to whole genomes will allow new or less researched genomes to be scanned quickly and annotated with miRNA data. The majority of the problems identified during this project can be solved relatively easily to produce a tool of greater quality.

9. Table of Figures

10. References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic Local Alignment Search Tool. *Journal of Molecular Biology , 215*, 403-410.

Artzi, S., Kiezun, A., & Shomron, N. (2008). miRNAminer: A tool for homologous microRNA gene search. *BMC Bioinformatics , 9* (39), 1471-2105.

Bentwich, I., Avniel, A., Karov, Y., Aharonov, R., Gilad, S., Barad, O., et al. (2005). Identification of hundreds of conserved and nonconserved human microRNAs. *Nature Genetics , 37* (7), 766 - 770.

Berezikov, E., Guryev, V., van de Belt, J., Wienholds, E., Plasterk, R. H., & Cuppen, E. (2005). Phylogenetic Shadowing and Computational Identification of Human microRNA Genes. *Cell* (120), 21–24.

Bernstein, E., Caudy, A. A., Hammond, S. M., & Hannon, G. J. (2001). Role for a bidentate ribonuclease in the initiation step of RNA interference. *Nature , 409*, 363-366.

Birney, E., Andrews, D., & Bevan, P. (2004). An Overview of Ensembl. *Genome Research , 14*, 925-928.

Bohnsack, M. T., Czaplinski, K., & Görlich, D. (2004). Exportin 5 is a RanGTP-dependent dsRNA-binding protein that mediates nuclear export of pre-miRNAs. *RNA* (10), 185-191.

Chendrimada, T. P., Finn, K. J., Ji, X., Baillat, D., Gregory, R. I., Liebhaber, S. A., et al. (2007). MicroRNA silencing through RISC recruitment of eIF6. *Nature , 447*, 823-828.

Clarke, N. J., & Sanseau, P. (2007). *microRNAs: Biology, Function & Expression.* Eagleville, PA 19408, USA: DNA Press, LLC.

Filipowicz, W. (2005). RNAi: The Nuts and Bolts of the RISC Machine. *Cell , 122*, 17–20.

Gregory, R. I., Chendrimada, T. P., Cooch, N., & Shiekhattar, R. (2005). Human RISC Couples MicroRNA Biogenesis and Posttranscriptional Gene Silencing. *Cell , 123*, 631–640.

Griffiths-Jones, S., Grocock, R., van Dongen, S., Bateman, A., & Enright, A. (2006). miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Research (Database Issue) , 34*, D140-D144.

Griffiths-Jones, S., Saini, H., van Dongen, S., & Enright, A. (2008). miRBase: tools for microRNA genomics. *Nucleic Acids Research (Database Issue) , 36*, D154-D158.

Hertel, J., & Stadler, P. F. (2006). Hairpins in a Haystack: recognizing microRNA precursors in comparative genomics data. *Bioinformatics , 22* (14), e197–e202.

Kim, J.-W., Kim, J., Kim, S.-K., & Zhang, B.-T. (2006). ProMiR II: a web server for clustered, nonclustered, conserved, nonconserved miRNA prediction. *Nucleic Acids Research (Webserver Issue) , 34*, W455-W458.

Lee, Y., Ahn, C., Han, J., Choi, H., Kim, J., Yim, J., et al. (2003). The nuclear RNase III Drosha initiates microRNA processing. *Nature , 425*, 415–419.

Lee, Y., Kim, M., Han, J., Yeom, K.-H., Lee, S., Baek, S. H., et al. (2004). MicroRNA genes are transcribed by RNA polymerase II. *The EMBO Journal , 23*, 4051-4060.

miRNA prediction with miR-abela. (n.d.). Retrieved from http://www.mirz.unibas.ch/cgi/pred_miRNA_genes.cgi

NCBI BLAST . (n.d.). Retrieved from http://www.ncbi.nlm.nih.gov/BLAST

RNA Secondary Structure Prediction and Comparison. (n.d.). Retrieved from http://www.tbi.univie.ac.at/RNA/

Sheng, Y., Engstrom, P. G., & Lenhard, B. (2007). Mammalian MicroRNA Prediction through a Support Vector Machine Model of Sequence and Structure. *PLoS ONE , 2* (9), e946.

Yi, R., Qin, Y., Macara, I. G., & Cullen, B. R. (2003). Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs. *Genes & Development , 17*, 3011-3016.

Zeng, Y., & Cullen, B. R. (2005). Efficient Processing of Primary microRNA Hairpins by Drosha Requires Flanking Nonstructured RNA Sequences. *The Journal of Biological Chemistry , 280* (30), 27595–27603.